# Soapy or Scrumptious: Cilantro Preference

Population structure analysis using SNPs

**BIO392**

Ayesha Gerber

Sarah Loose

Universität Zürich[UZH]

# Intro

- **What do you associate Cilantro taste with?**

- **Heritable Hate:**
  -> twin studies show dislike might depend on genetic variations and predispositions

- **Gene Candidates determining cilantro dislike:**

| *OR6A2* | *TAS2R1* | *OR4N5* |
|---|---|---|
| Receptor, highly sensitive to aldehyde chemicals | Bitterness receptor | Odorant receptor |

23andMe

UNIVERSITY OF TORONTO

# Data & Project Goal

- Biallelic SNVs Data (VCF format) from the 1000 Genome Project, published 2018
- **Can we find preference patterns in cilantro preference based on ancestry inferred from SNVs?**

# Analysis Workflow

**VCF file --> PLINK convert --> LD Pruning --> PCA analysis -- > ADMIXTURE Analysis**

- **PLINK convert:** filter our genomic region of interest and convert .vcf to .bed file (Data processing)
- **Linkage Disequilibrium Pruning:** remove SNPs with high linkage disequilibrium to prevent bias and remove redundancy (Data processing)
- **PCA:** linear orthogonal transformation reducing dimensionality retaining components explaining the greatest variance (Visualization)
- **ADMIXTURE:** maximum likelihood method to infer ancestry proportions of an individual (Anscestry Inference)

# LD Pruning – Why and How?

**Nonrandom association of alleles at multiple DNA markers that results from their close proximity to one another within a chromosome and co-inheritance.**

- make the analysis more robust, avoiding bias by redundant information
- We give 3 values into the command line

plink2 --bfile chr11_fix \ --indep-pairwise **50 10 0.1** \ --out ld_chr11

50: window size of SNPs

10: step of frame shift

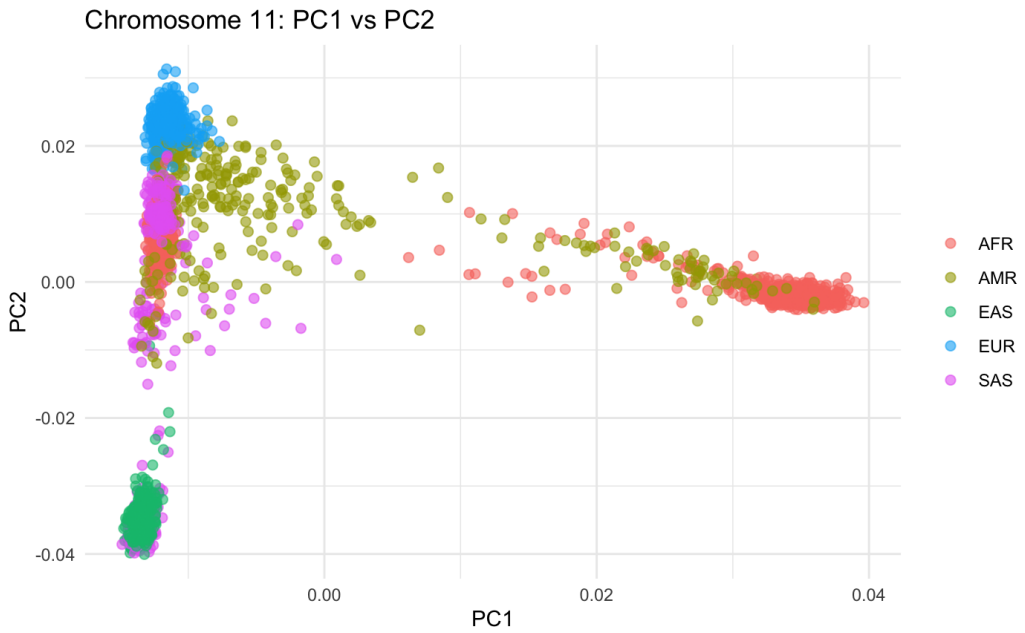
0.1: $r^2$ threshold tolerated (how correlated are the SNPs, pairwise correlation between 2 SNPs in the sliding window)

# PCA - Making Complex Data Simpler

- Every person has thousands of SNPs
- PCA reduces this huge amount of information into just a few variables

  -> **principal components**

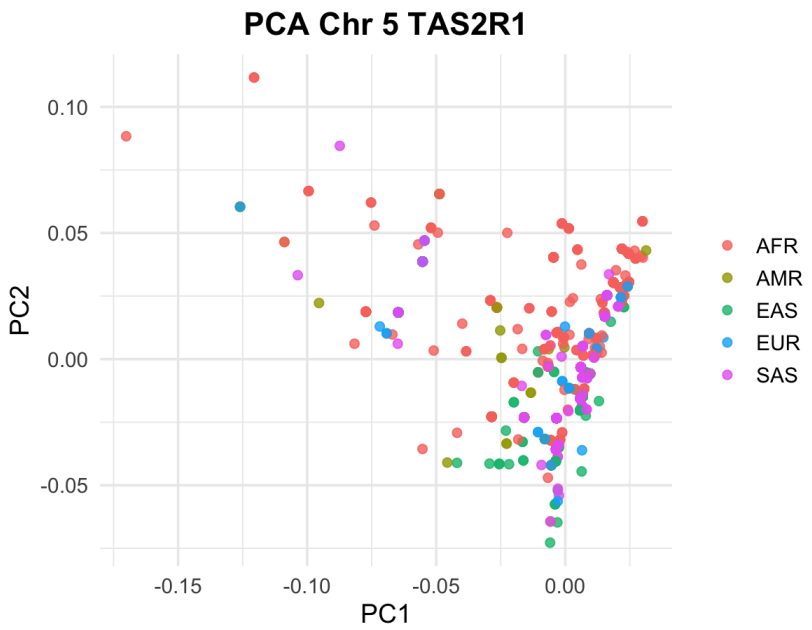
- It allows us to plot people as points, based on how genetically similar they are.

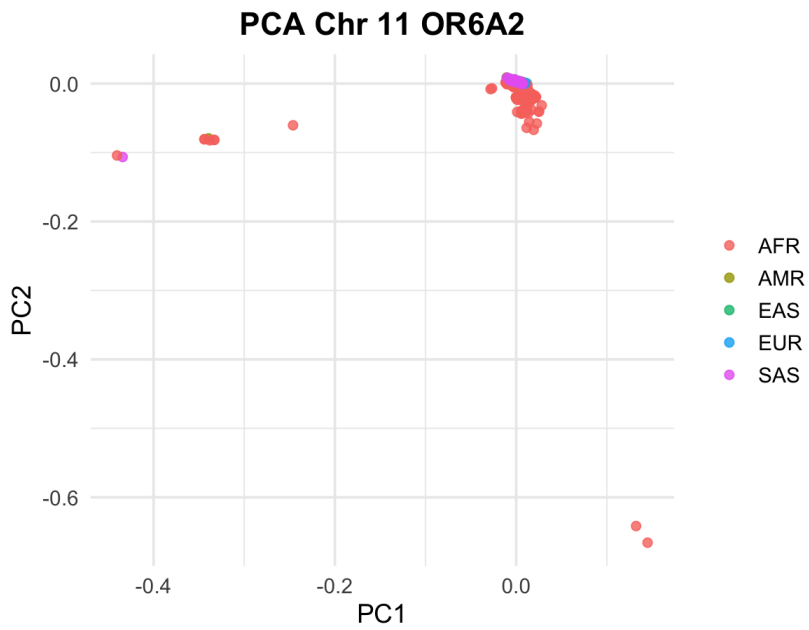


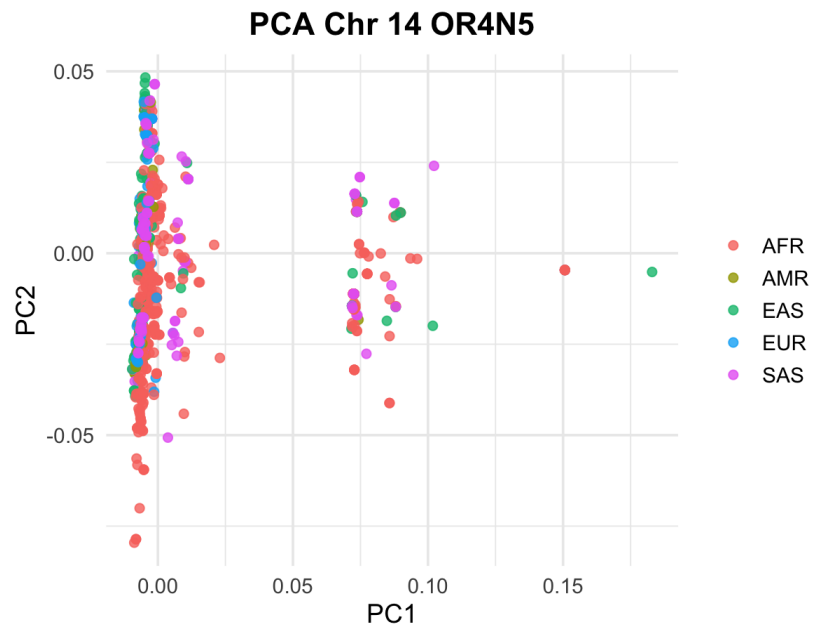

Whole chromosome 11.
We see SNP variance between the superpopulations.
Nothing inferable about cilantro.

➢ We used PCA to see if people with different ancestry also differ in the **genes linked to cilantro taste**.

# Results: PCA



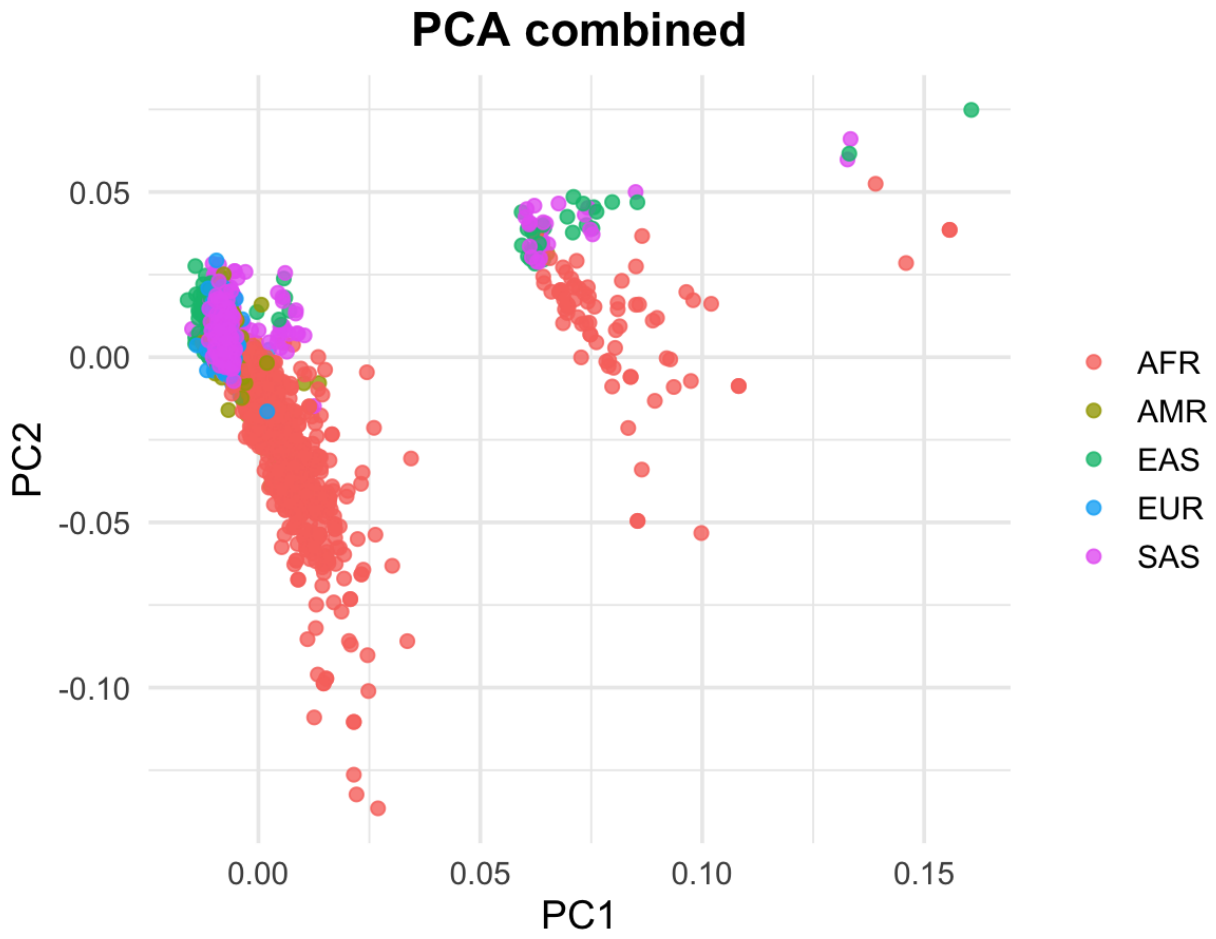




- Some minor clustering tendencies, but overall overlap between populations

➢ Low number of SNPs  leads to weak differentiation and no clear structure.
➢ Functional SNPs may exist, but they do not create population-wide structure.

# Results: PCA




- Three loose groupings
  - ➢ Could reflect gene differences
- Within each cluster, population structure is slightly reflected:
  - EUR, SAS, and EAS tend to cluster together.
  - AFR individuals appear more spread out.

➢ pattern could reflect regional variation across the three loci

➢ overall structure is still weak

# ADMIXTURE – Infer Population Structure

**Input:**

**SNP Data** of all individuals

**K** = number of ancestral populations the model will infer

**Process:**

**Maximum likelihood estimation** to:

- Determine each individuals **proportion of the genome** from each of the K **ancestral populations**
- Match **allele frequency** for each SNP to one K ancestral population
- Try **several runs** to find best K (using **cross-validation**)

**Output:**

**.Q file** (ancestry proportions per individual) -> visualization

**.P file** (allele frequencies per population)

# ADMIXTURE RUN

```
(base) sarah@MacBook-Air-von-Sarah admixture_macosx-1.3.0 % ./admixture
-s 12345  /Users/sarah/Desktop/day10/chr11_gefiltert.bed 6
****                  ADMIXTURE Version 1.3.0                 ****
****                  Copyright 2008-2015                     ****
****           David Alexander, Suyash Shringarpure,          ****
****                John  Novembre, Ken Lange                 ****
****                                                          ****
****                Please cite our paper!                    ****
****   Information at www.genetics.ucla.edu/software/admixture  ****

Random seed: 12345
Point estimation method: Block relaxation algorithm
Convergence acceleration algorithm: QuasiNewton, 3 secant conditions
Point estimation will terminate when objective function delta < 0.0001
Estimation of standard errors disabled; will compute point estimates on
y.
Size of G: 2548x242
Performing five EM steps to prime main algorithm
1 (EM)  Elapsed: 0.885  Loglikelihood: -36795.8 (delta): 885978
2 (EM)  Elapsed: 0.885  Loglikelihood: -34320.4 (delta): 2475.4
3 (EM)  Elapsed: 0.891  Loglikelihood: -34132.2 (delta): 188.224
4 (EM)  Elapsed: 0.9    Loglikelihood: -33982.3 (delta): 149.88
5 (EM)  Elapsed: 0.9    Loglikelihood: -33857.1 (delta): 125.213
Initial loglikelihood: -33857.1
Starting main algorithm
1 (QN/Block)    Elapsed: 0.563  Loglikelihood: -25819.8 (delta): 8037.3
2 (QN/Block)    Elapsed: 0.597  Loglikelihood: -21984.8 (delta): 3834.9
3 (QN/Block)    Elapsed: 1.027  Loglikelihood: -20367.5 (delta): 1617.3
4 (QN/Block)    Elapsed: 0.884  Loglikelihood: -19498.4 (delta): 869.03
5 (QN/Block)    Elapsed: 1.024  Loglikelihood: -19015.9 (delta): 482.47
6 (QN/Block)    Elapsed: 1.335  Loglikelihood: -18884.4 (delta): 131.54
7 (QN/Block)    Elapsed: 0.888  Loglikelihood: -18856   (delta): 28.438
8 (QN/Block)    Elapsed: 0.887  Loglikelihood: -18829.5 (delta): 26.498
9 (QN/Block)    Elapsed: 1.175  Loglikelihood: -18819.2 (delta): 10.217
10 (QN/Block)   Elapsed: 1.324  Loglikelihood: -18818.9 (delta): 0.3655
3
11 (QN/Block)   Elapsed: 0.888  Loglikelihood: -18818.9 (delta): 4.9003
e-06
Summary:
Converged in 11 iterations (17.612 sec)
Loglikelihood: -18818.881469
Fst divergences between estimated populations:
        Pop0    Pop1    Pop2    Pop3    Pop4
Pop0
Pop1    0.470
Pop2    0.834   0.470
Pop3    0.871   0.468   0.831
Pop4    0.560   0.301   0.554   0.559
Pop5    0.904   0.554   0.769   0.902   0.623
Writing output files.
```

Set arbitrary **K (an. pop.)**

Set **random seed** for reproducibility

**Expectation Maximization:**

Iterative, find biggest Log-likelihood fast

**Quasi-Newton Algorithm**

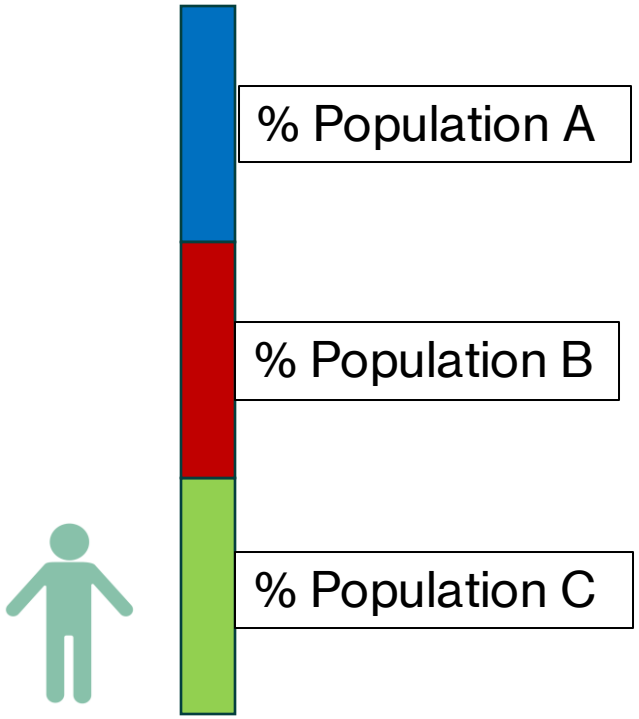
Iterative Log-likelihood optimization

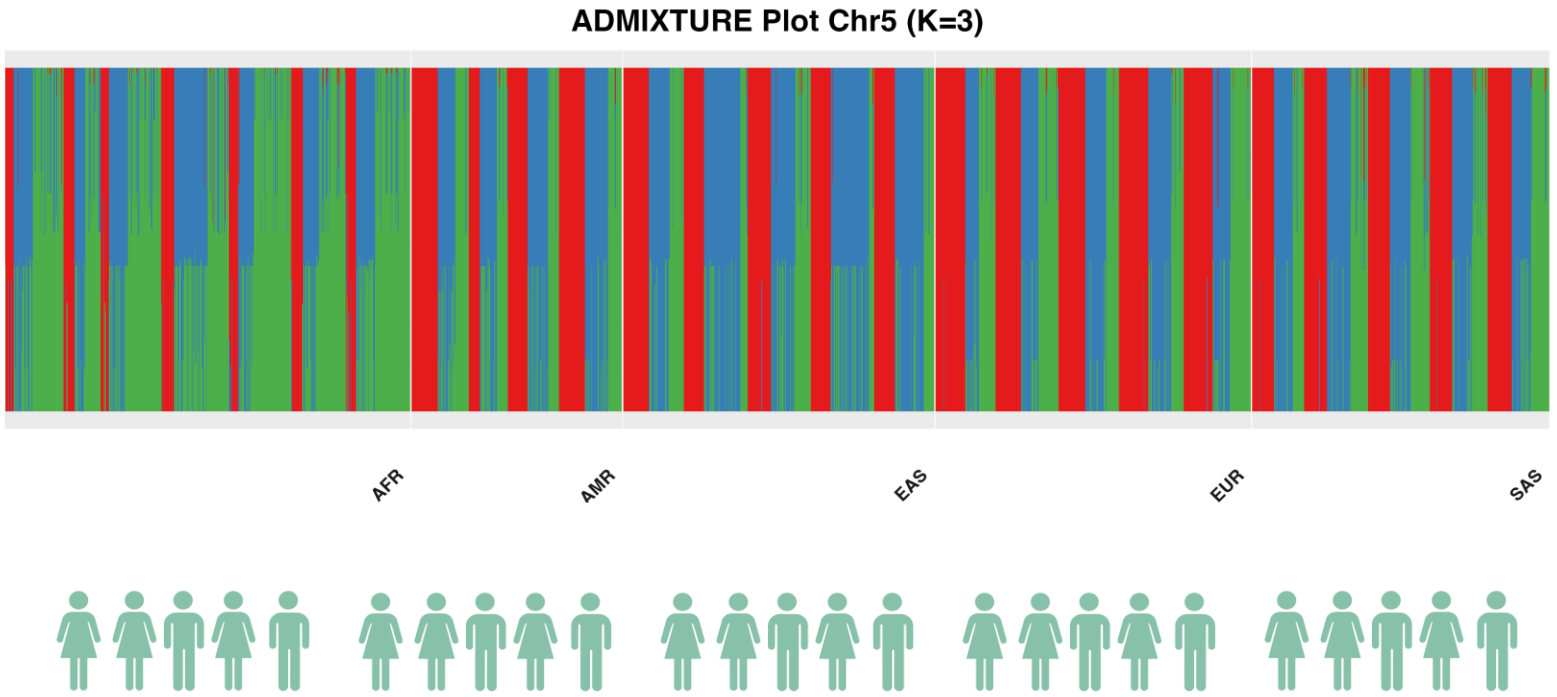
**Fst Divergence Matrix**

How different are the K populations based on allele frequency

Write .**Q** and .**P** files

# Visualization from .Q file



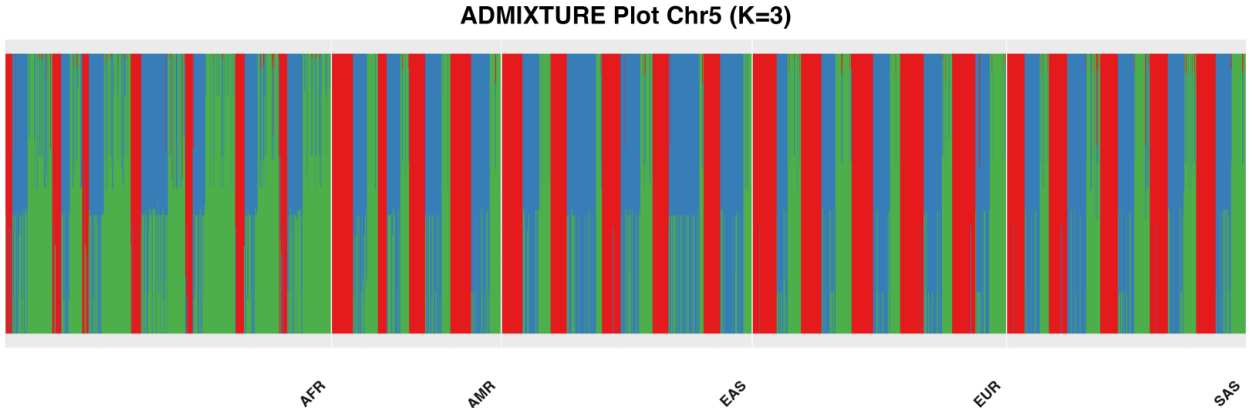



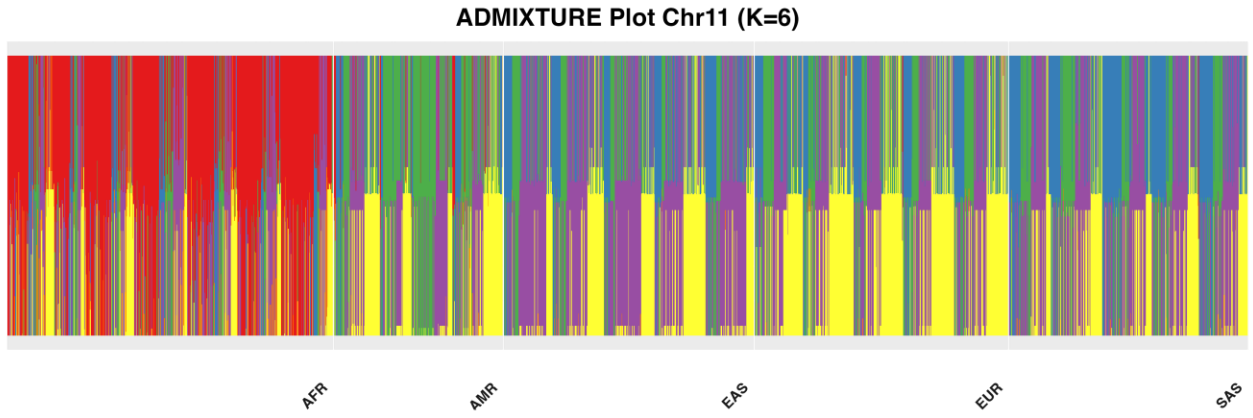
# Results: ADMIXTURE

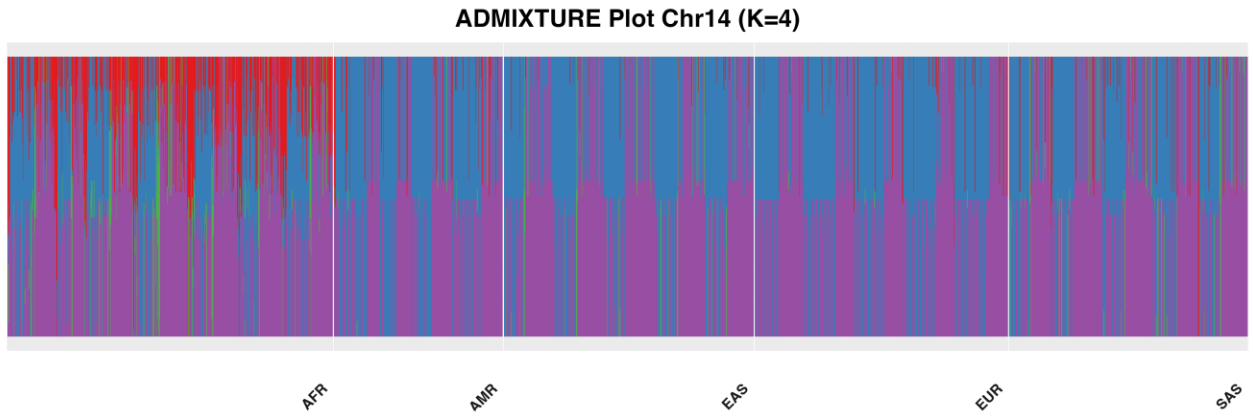
- highly mixed ancestry components across all populations.
- No clear population-specific clustering is visible.

➢ Suggests low between-population differentiation at all loci.

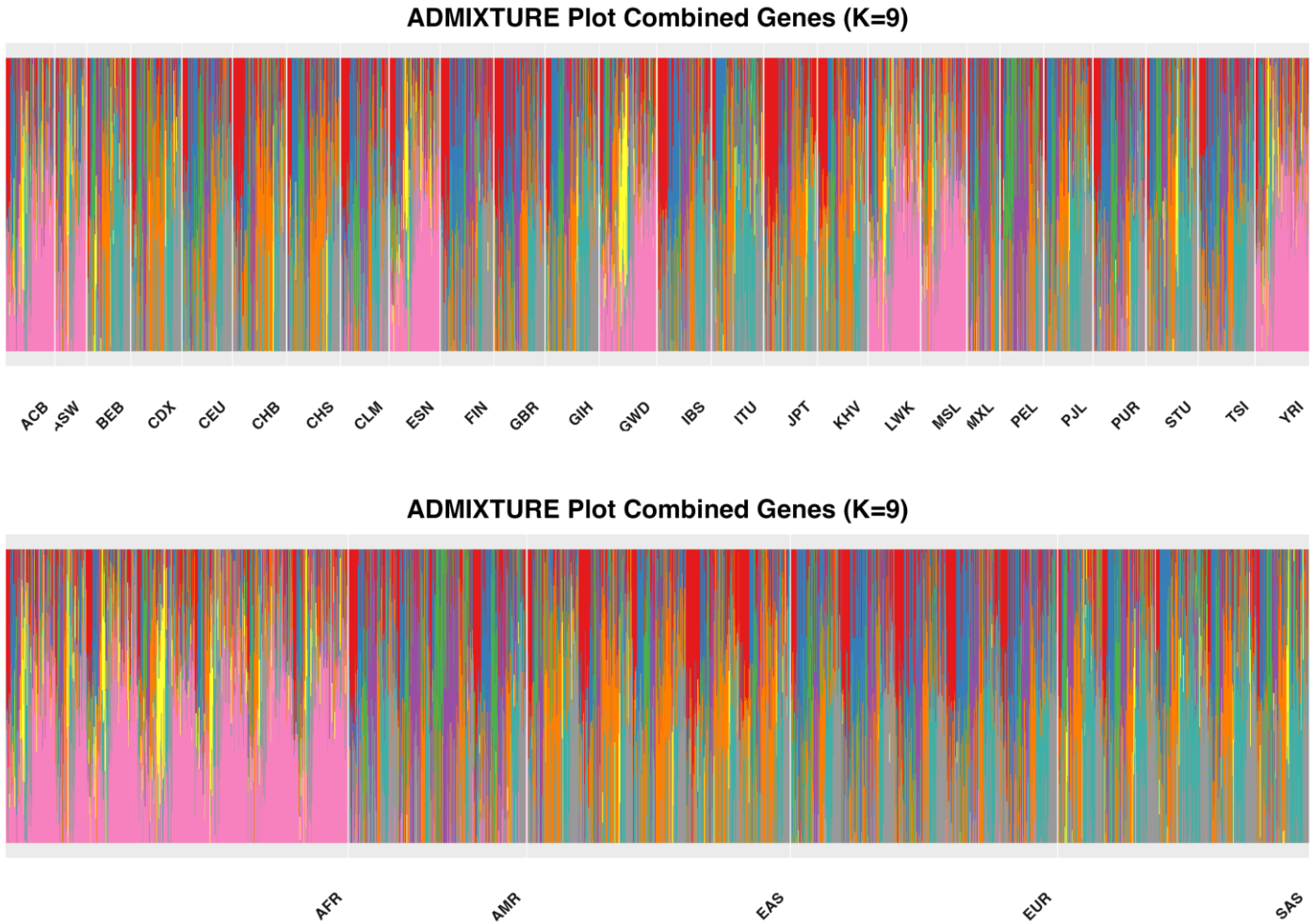
➢ Too few SNPs and insufficient variation in this region to detect population structure.



# Results: ADMIXTURE




- still shows no clear ancestry patterns or continental clusters
- All populations are highly admixed across clusters

➢ Combining genes increased SNP number, but not enough to overcome the noise

# Discussion

- The **combined PCA** shows some subtle grouping, but **not enough** for strong conclusions.
- **None** of the individual genes or the combined dataset showed **clear ancestry patterns** in ADMIXTURE.

➢ PCA & ADMIXTURE is **better suited for genome-wide data or large SNP sets**.

# Outlook

- **ADMIXTURE** and **PCA**
  - Enables exploration of genetic variation between populations by visualizing ancestry proportions
  - Applications in predicting population distribution of disease markers

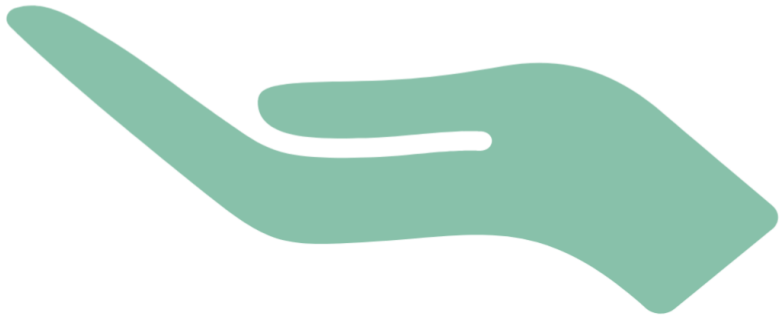
➢ are powerful tools, but only when used with **large-scale, genome-wide data**.

**For small genomic regions** or specific traits like cilantro perception, other methods are more appropriate, such as:

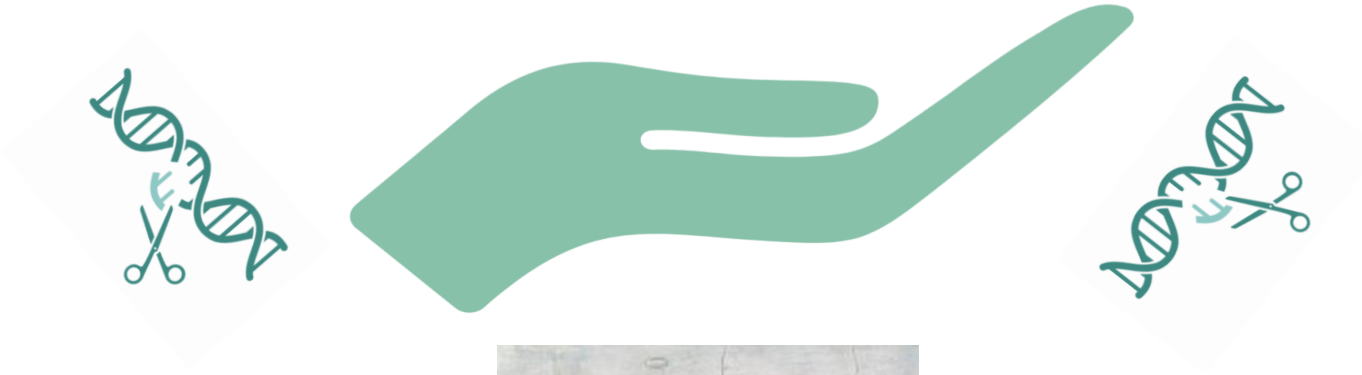
- Genome-wide association studies (**GWAS)**
- **Allele frequency comparisons**

- Olfactory receptors are highly variable in the human population (SNPs, Copy number, Pseudogenes) but have high sequence similarity among each other.
- This could lead to biases against SNPs in these regions due to filtering and sequencing methods and misrepresentation of the regions in the 1000 Genome project.

# If you don´t like cilantro try:



AGAIN



# Sources


Liu, CC., Shringarpure, S., Lange, K., Novembre, J. (2020). Exploring Population Structure with Admixture Models and Principal Component Analysis. In: Dutheil, J.Y. (eds) Statistical Population Genomics. Methods in Molecular Biology, vol 2090. Humana, New York, NY. https:doi.org/10.1007/978-1-0716-0199-0_4

https://www.23andme.com/en-int/genetic-science/ Accessed 01.05.2025

https://www.uniprot.org/uniprotkb/Q9NYW7/entry Accessed 01.05.2025

https://www.uniprot.org/uniprotkb/Q8IXE1/entry#function Accessed 01.05.2025

Eriksson, N., Wu, S., Do, C.B. *et al*. A genetic variant near olfactory receptor genes influences cilantro preference. *Flavour* **1**, 22 (2012). https://doi.org/10.1186/2044-7248-1-22

Mauer, L. K. Genetic Determinants of Cilantro Preference. MSc thesis, Univ. Toronto https://tspace.library.utoronto.ca/bitstream/1807/31335/1/Mauer_Lilli_K_201108_MSc_Thesis.pdf (2011).

Callaway, E. Soapy taste of coriander linked to genetic variants. *Nature* (2012). https://doi.org/10.1038/nature.2012.11398

**PCA plots were created in R. Population Distribution plots were created with Admixture. Cartoons were created in Biorender.**

C. Trimmer et. Al. Genetic variation across the human olfactory receptor repertoire alters odor perception, *Proc. Natl. Acad. Sci. U.S.A.* 116 (19) 9475-9480, https://doi.org/10.1073/pnas.1804106115 (2019).

# **Thesis** Genetic Determinants of Cilantro Preference

- **Toronto Nutrigenomics and Health Study** Participants (n = 1,639; 1,117 women and 522 men) which is a cross-sectional study investigating gene-diet interactions and biomarkers of chronic disease, as well as genetic determinants of eating behaviors.
- **Subjects** were between 20 and 29 years of age
- **Excluded:** pregnant or breastfeeding, non-english speakers, or who did not provide a 12- hour fasting venous blood sample, Smokers (n = 105), Subjects with any missing data (n = 10),Subjects who listed more than one ethnicity (n=143) or any group with fewer than 20 subjects were excluded from the current analyses
- **ethnocultural groups (**based in self-identification):
  six groups (**Caucasian,** n = 581; **East Asian**, n = 540; **South Asian**, n = 165; **Middle Eastern,** n =36; **African descent**, n = 32; and **Hispanic,** n = 27).
- After exclusions, the final sample population consisted of 1,381 subjects (962 women and 419 men).
- **East Asians and Caucasians had the highest prevalence of cilantro dislikers**
- **genome-wide scans were performed using an Affymetrix 6.0 chip. A totalof 16 SNPs reached GWAS significance (p<5.5x10-6).**
- SNPs: OR4N5 chr14 and TAS5R2 chr 5
- **75% of individuals homozygous for the minor allele of both SNPs reported disliking, whereas 0% of subjects homozygous for the major allele of both SNPs reported disliking**

# **Thesis** Genetic Determinants of Cilantro Preference

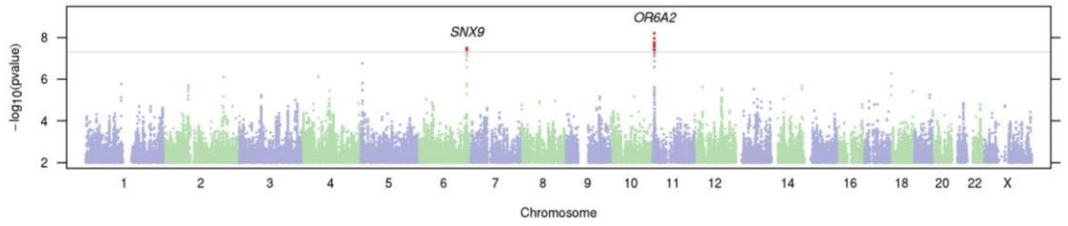
- At the SNP level, markers eliminated first were those with call rates less than 95% (16,781 SNPs). Hardy-Weinberg equilibrium (HWE) was then assessed, and SNPs (30,711) with HWE P values less than 1×10-8 were excluded, as this suggests that they are not in HWE in this population.
- **Hardy-Weinberg Equilibrium (HWE) Filter: genotype frequencies should follow expected patterns** if the population is not evolving (no selection, mating is random, etc.).

# A genetic variant near olfactory receptor genes influences cilantro preference

- SNP, rs72921001 ($p = 6.4 \times 10^{-9}$, odds ratio 0.81 per A allele), lies within a cluster of olfactory receptor genes on chromosome 11. Among these olfactory receptor genes is ***OR6A2*,** which has a high binding specificity for several of the aldehydes that give cilantro its characteristic odor. We also estimate the **heritability** of cilantro soapy-taste detection in our cohort, showing that the heritability t**agged by common SNPs is low, about 0.087.**

**In both the GWAS set and the replication set, all participants were of European ancestry.**

- On the 23andMe website, participants contribute information through a combination of **research surveys** (longer, more formal questionnaires) and research 'snippets' (multiple-choice questions appearing as part of various 23andMe webpages). In this study, participants were asked two questions about cilantro via research snippets:
- 'Does fresh cilantro taste like soap to you?' (Yes/No/I'm not sure)
- 'Do you like the taste of fresh (not dried) cilantro?' (Yes/No/I'm not sure)
- Subjects were **genotyped** on one or more of three chips, two based on the Illumina HumanHap550+ BeadChip and the third based on the Illumina OmniExpress+ BeadChip (San Diego, CA, USA). The platforms contained 586,916, 584,942, and 1,008,948 SNPs. Totals of 291, 5,394, and 10,184 participants (

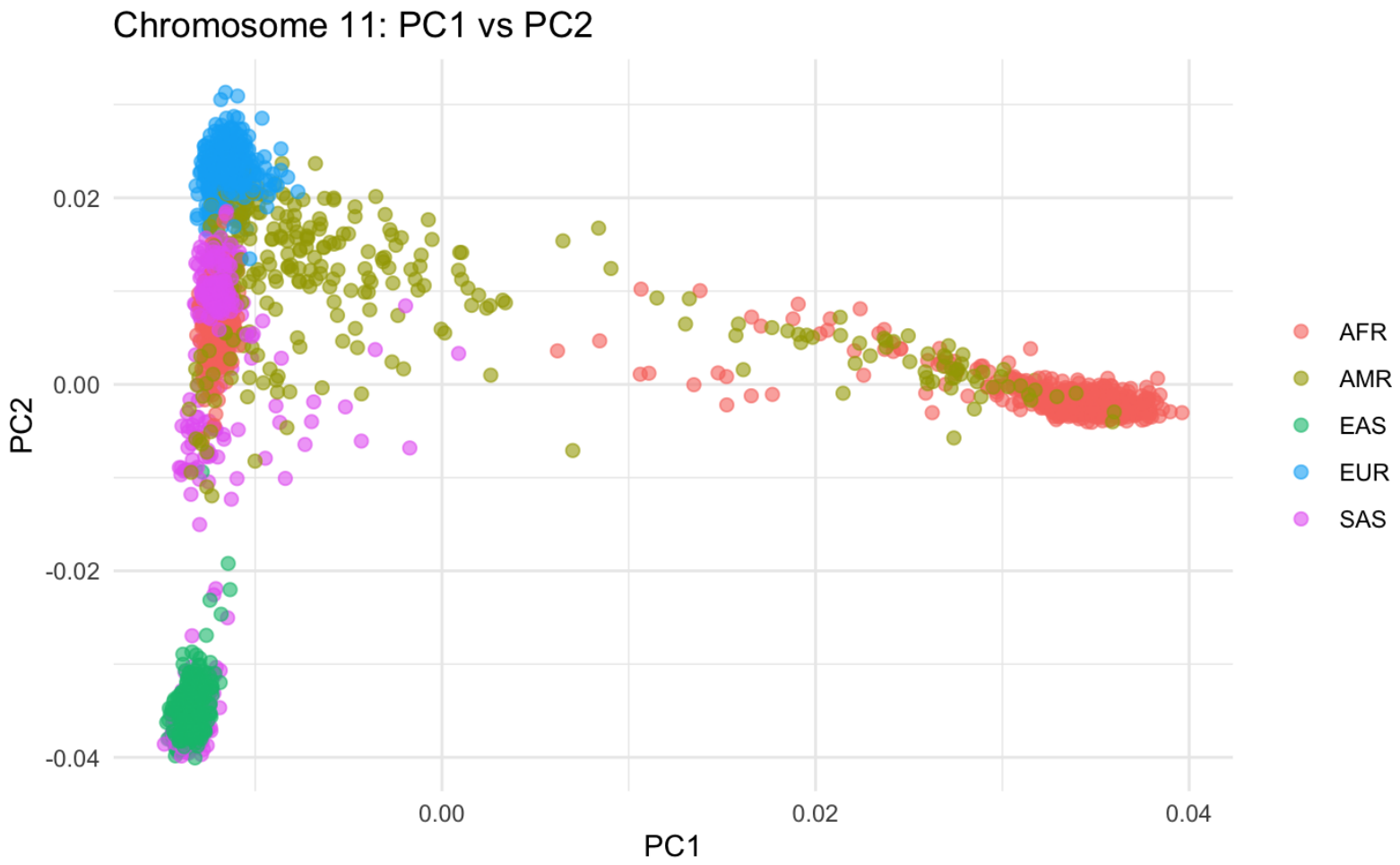



**Manhattan plot of association with cilantro soapy-taste.** Negative $\log_{10} p$ values across all SNPs tested. SNPs shown in *red* are genome-wide significant ($p < 5 \times 10^{-8}$). Regions are named with the postulated candidate gene.

# 23andMe

- Autosomal analysis, genotyping based on SNPs
- Samples that yield sufficient quantities of DNA are **genotyped** on our custom SNP chip (or **microarray)**.
- This process is performed in batches of approximately 96 samples. We monitor the process to make sure it is going as expected.
- -> so NO WES or WGS, only SNPs of interest in array-based methods

# Whole Chromosome 11 PCA




- Whole Chromosome
- Nothing inferable about cilantro anymore
- We see SNP variance between the superpopulations

# GWAS Workflow

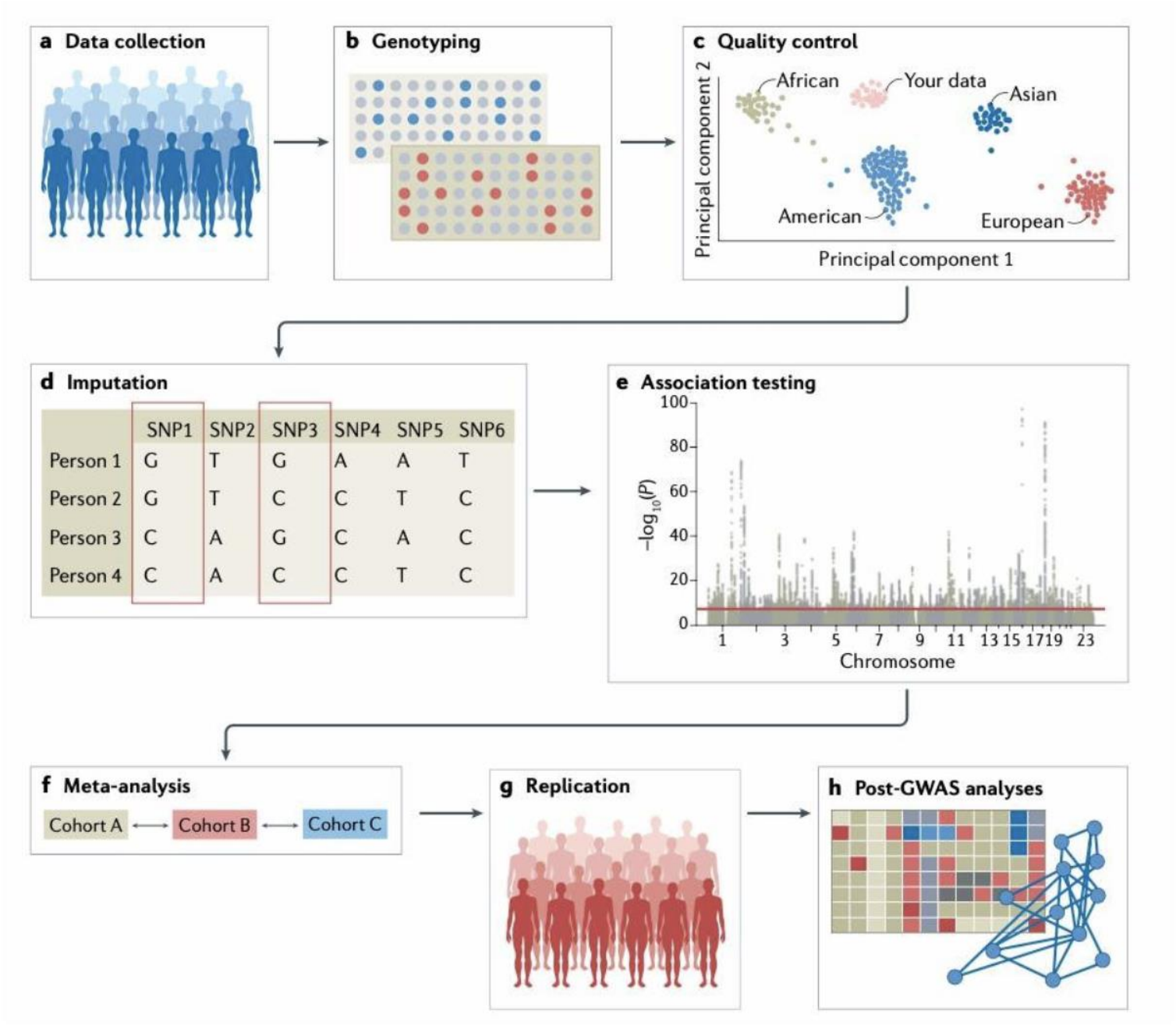



Figure 1.Overview of steps for conducting GWAS.(Uffelmann, et.al, 2021)

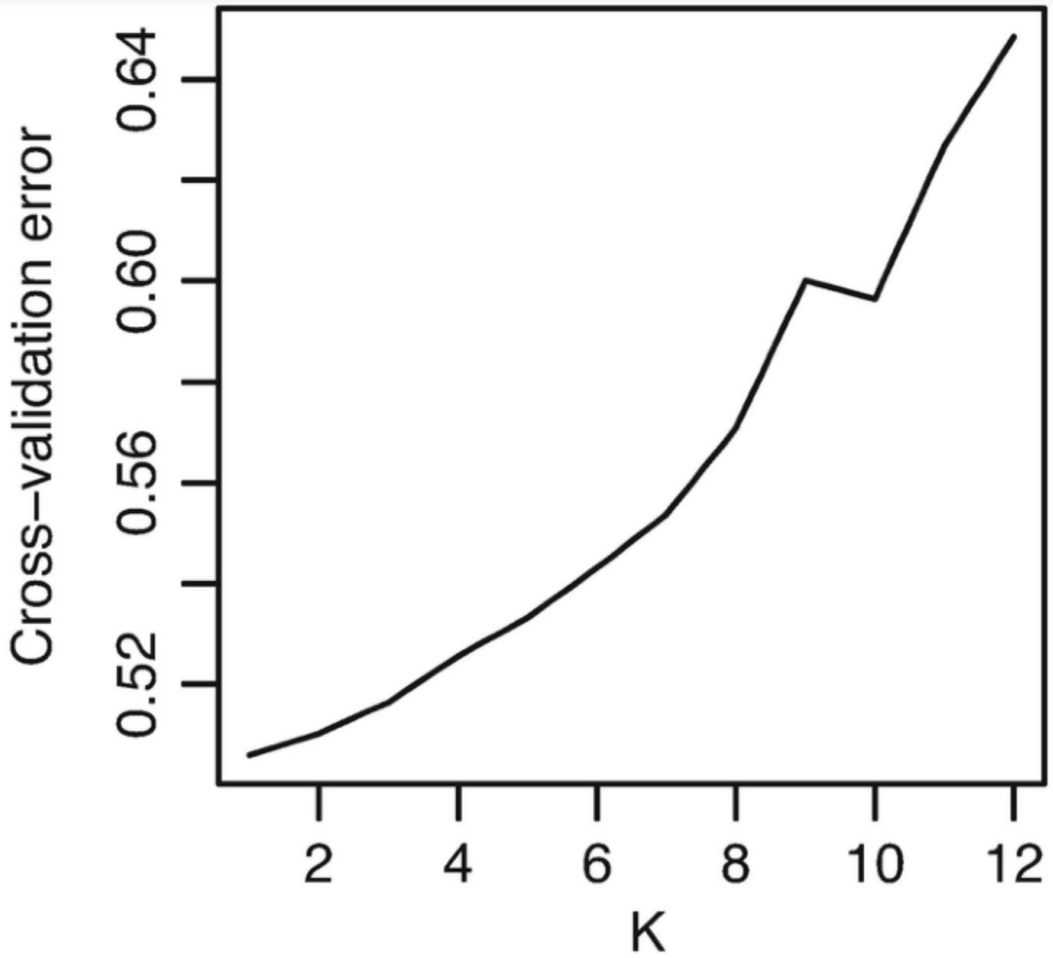
https://www.cd-genomics.com/resource-gwas-vs-whole-genome-sequencing.html

02.05.2025

# Cross-Validation

- How to choose the perfect K (ancestral populations)
- Accurate fitting without overfitting
- computed as the **negative log-likelihood**
- Smaller error = better fit
- But also visually inspect the barplots to decide on perfect K

"The selection of *K* is a difficult problem to automate in a way that is robust."




cross-validation error suggests a **single** source population can model the data adequately and larger values of *K* lead to over-fitting